Multiple Instance Metric Learning from Automatically Labeled Bags of Faces

Matthieu Guillaumin, Jakob Verbeek, Cordelia Schmid

LEAR, INRIA Grenoble, LJK, France

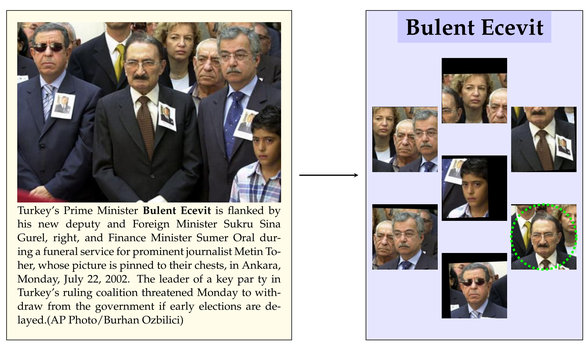
Summary

- Weakly supervised learning of Mahalanobis metrics for unconstrained face recognition using images with captions.
- Approach: multi-instance multi-label metric learning from bag-level labels.
- We compare with a "missing label" approach where instance labels are iteratively inferred.
- We introduce the *Labeled Yahoo! News* data set, with captions and manual annotations.
- We show that MildML metrics outperform similar metrics from automatically inferred instance-level labels for face recognition,
- With clean bag labels, MildML metrics perform comparably to metrics learned from instance-level labels.

Motivation of our work

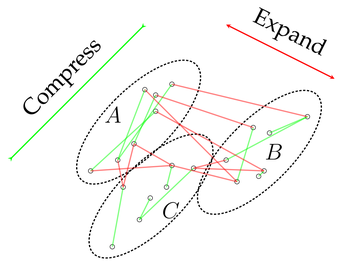
Manually annotating the tens of thousands of training samples required for robust metric learning is a **costly process**. Is it possible to reduce the amount of annotation effort, and even **remove any user intervention** ?

Bags of faces from news images



Labels can be obtained automatically from news images with captions, using NLP and face detection. However, those labels are imperfect, noisy and are at bag level.

Supervised setting for metric learning



Data $\mathbf{x}_i$ is manually labeled with y_i. Learn a Mahalanobis distance d_{M} that makes images of positive pairs (same class) closer than those of negative pairs (different class):

$$d_{\mathrm{M}}(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i - \mathbf{x}_j)^\top \mathbf{M} (\mathbf{x}_i - \mathbf{x}_j), \qquad (1)$$

where M is a symmetric semi-definite positive matrix. ML can be kernelized and regularized to low-rank M.

Logistic Discriminant Metric Learning

In LDML [1], the probability p_{ij} that a pair is positive is modeled with the sigmoid function $\sigma(z) = (1+\exp(-z))^{-1}$:

$$p_{ij} = p(y_i^\top y_j = 1 | \mathbf{x}_i, \mathbf{x}_j; \mathbf{M}, b) = \sigma(b - d_{\mathrm{M}}(\mathbf{x}_i, \mathbf{x}_j)), \qquad (2)$$

where b is a bias term. Maximum likelihood estimation of M and b is performed using (projected) gradient ascent.

"Missing label" approach for LDML

When the labels y_i of instances are unknown, we extend the optimization problem to inferring them from the bag-level ones. The joint optimization is intractable. We optimize the objective by alternating between:

(1) Find M and b for fixed labels $\{y_i\}$: this is LDML.

(2) Find labels $\{y_i\}$ for fixed M and b: this problem is a constrained clustering known as *face naming* [2]:

$$\max_{\{y_i\}} \quad \sum_{i,j} w_{ij} y_i^\top y_j, \qquad (3)$$

where $w_{ij} = b - d_{\mathrm{M}}(\mathbf{x}_i, \mathbf{x}_j)$.

Contraints on $\{y_i\}$ are the following:

(i) Faces can only be assigned to names in the caption.

(ii) Faces can only be assigned to at most one name.

(iii) Names can only be assigned to at most one face.

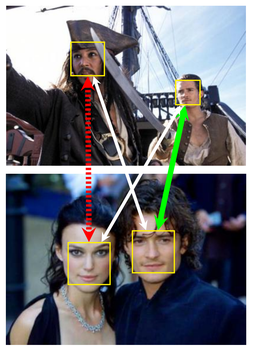
Local optimisation approach of [3] for graph-based clustering: iterate over bags while enforcing constraints until convergence.

Multiple instance LDML: MildML

A Multiple Instance Learning [4] point-of-view is also possible for metric learning. Images are seen as bags of instances $\mathcal{X}_d = \{\mathbf{x}_1^d, \mathbf{x}_2^d, ..., \mathbf{x}_{N_d}^d\}$.
The distance measure is extended to bags [5]:

$$d_{\mathrm{M}}(\mathcal{X}_d, \mathcal{X}_e) = \min_{1 \le k \le N_d, 1 \le l \le N_e} d_{\mathrm{M}}(\mathbf{x}_k^d, \mathbf{x}_l^e). \qquad (4)$$

This can be seen as a **robust selection of a pair** among all face pairs between bags.



Johnny Depp, Orlando Bloom.

Gore Verbinski, Jerry Bruckheimer, Johnny Depp, Keira Knightley, Orlando Bloom.

Intersection of bag-level labels is used to define positive ($t_{de} = 1$) and negative ($t_{de} = 0$) pairs of bags, and a LDML objective is used, with similar optimization:

$$\max_{\mathbf{M}, b} \quad \sum_{d,e} t_{de} \log p_{de} + (1 - t_{de}) \log(1 - p_{de}). \qquad (5)$$

Data set, features and tasks

Labeled Yahoo! News

Contains around 30000 news images with captions. We split it in two independent parts (nobody appears in both) for training and testing. Manually annotated, available online: `http://lear.inrialpes.fr/data/`

Features

Features are extracted at 9 locations on the face using SIFT descriptors at 3 different scales: 3456D face descriptor.

Tasks

(I) Face verification: decide for image pairs if they depict the same person.

(II) Face naming: name all faces in the data set.

Experiments

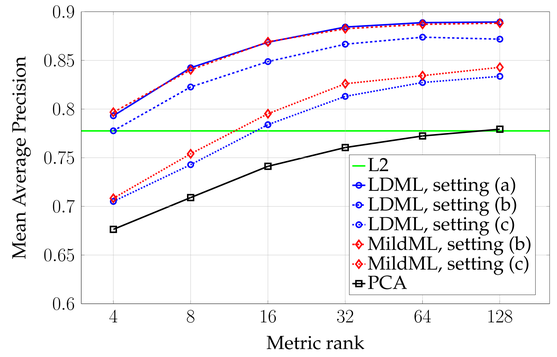
Three settings with varying amount of supervision:

(a) Clean instance labels (supervised ML)

(b) Clean bag labels

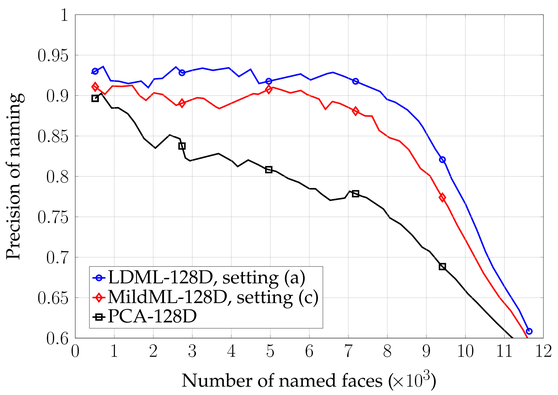
(c) Noisy bag labels (fully automatic)

Face verification



- MildML (b) performs comparably to LDML (a).
- MildML outperforms LDML for any given setting.
- MildML (c) closer to LDML (a) than to PCA.

Face naming



- Fully automatic metric close to supervised one.

References

[1] M. Guillaumin, J. Verbeek and C. Schmid. Is that you? Metric learning approaches for face recognition. ICCV, 2009.

[2] T. Berg, A. Berg, J. Edwards, M. Maire, R. White, Y. Teh, E. Learned-Miller, D. Forsyth. Names and faces in the news. CVPR, 2004.

[3] M. Guillaumin, T. Mensink, J. Verbeek and C. Schmid. Automatic face naming using caption-based supervision. CVPR, 2008.

[4] T. Dietterich, R. Lathrop, T. Lozano-Perez, A. Pharmaceutical. Solving the multi-instance problem with axis-parallel rectangles. AI, 1997.

[5] R. Jin, S. Wang, Z.H. Zhou. Learning a distance metric from multi-instance multi-label data. CVPR, 2009.